ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

COMPARISON OF FILTER BASED FEATURE SELECTION ALGORITHMS: AN OVERVIEW

R. Porkodi, Assistant Professor, Department of Computer Science, Bharathiar University, India, porkodi_r76@yahoo.co.in

ABSTRACT

Feature selection is very much useful to choose a subset of features from data set containing more than 100 to 1000 attributes by eliminating irrelevant features to improve predictive information. Feature selection is the most promising field of research in data mining in which most impressive achievements have been reported. The feature selection influences the predictive accuracy of any data set. Hence, it is essential to study the metrics that are already used in this area. This paper provides the clear insight to different feature selection methods reported in the literature and also compares all methods with each other. The experimental result shows that the feature selection methods provide better result for breast cancer data set.

KEYWORDS: Filter model, Wrapper model, Clustering, Classification, Feature selection, Accuracy

I. INTRODUCTION

Data mining an interdisciplinary subfield of computer science which is the computational process of discovering patterns in large data sets by intersection of methods such as artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining involves six common tasks such as Anomaly detection , Association rule mining, Clustering, Classification, Regression and Summarization. The rapid growth and advancements in knowledge datasets and computer techniques motivated the data accumulation in high speed. The all above said tasks in data mining requires the knowledge datasets to be processed to obtain any sort of understandable structure.

The processing of accumulated data itself has become a big challenge for researchers in order to identify relevant and irrelevant features to improve the predictive accuracy and for this the number of data reduction techniques has been proposed so far. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance [1]. Feature selection is one of the important and frequently used techniques in data reduction or preprocessing for data mining. There are a number of advantages of feature selection includes it reduces the number of features, removes irrelevant, redundant, or noisy data, reduce the computational cost, speeding up a data mining algorithm and improve the classification accuracy [2].

Feature selection is a process that selects a subset of original features. The optimality of a feature subset is measured by an evaluation criterion. The feature selection process consists of four basic steps as shown in Figure 1, namely, subset generation, subset evaluation, stopping criterion, and result validation [3]. Subset generation is a search procedure [4] that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Finally, the selected best subset to be validated by domain experts or any other test and the selected best subset may be given as an input to any data mining task.

The feature selection methods broadly classified into three categories: the *filter* model [5, 6, 7], the *wrapper* model [8, 9, 10], and the *hybrid* model [11, 12, 13]. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model [14]. The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

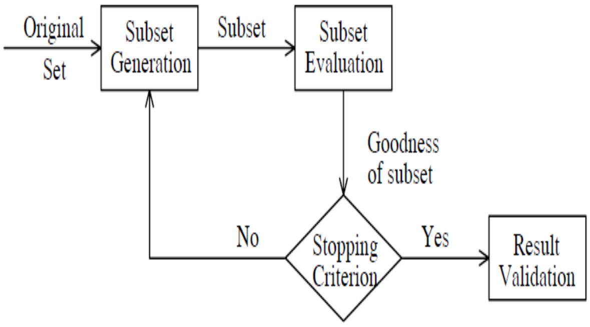

Fig 1 Feature Selection Process

This paper is organized as follows. Section 1 gives the overview and introduction of feature selection, section 2 discusses the review of literatures and section 3 discusses the different feature selection methods used in this paper. The experimental results are shown and discussed in section 4 and finally the paper is concluded in section 5.

II. RELATED WORK

Feature selection is one of the active fields of research for decades in machine learning, data mining, genomic analysis [15], text mining [16], image retrieval [17], intrusion detection [18], etc.

The paper [19] adopted an unbiased protocol to perform a fair comparison of frequently used multivariate and univariate gene selection techniques, in combination with a range of classifiers. In their conclusion they found that univariate and multivariate feature selection algorithms greatly improved the performance of cancer genes.

The authors [20] used naive bayes classifier with feature selection for medical data mining. Our experimental results indicate that, on an average, with the proposed CHI-WSS algorithm utilizing naïve Minimum Description Length (MDL) discretization, Chi-square feature selection ranking and wrapper approach, provides on the average better accuracy performance and feature dimensionality reduction.

The paper [21] evaluated several inter-class as well as probabilistic distance-based feature selection methods as to their effectiveness in preprocessing input data for inducing decision trees. They used real-world data to evaluate these feature selection methods. Results from this study show that inter-class distance measures result in better performance compared to probabilistic measures, in general

Authors in paper [22] proposed algorithm for feature selection is based on an application of a rough set method to the result of principal components analysis (PCA) used for feature projection and reduction. Finally, the paper presents numerical results of face and mammogram recognition experiments using neural network, with feature selection based on proposed PCA and rough set methods.

III. FEATURE SELECTION ALGORITHMS

This part of this paper briefly introduces the feature selection algorithms that has been discovered and reported in the literatures. The feature selection algorithms are classified into three categories such as filter model, wrapper model and embedded model according to the computational models. The filter model relies on the general characteristics of data and evaluates features without involving any learning algorithm. The wrapper model requires having a predetermined learning algorithm and uses its performance as evaluation criterion to select features. The embedded model incorporate variable selection as a part of the training process, and feature relevance is obtained analytically from the objective of the learning model.

A. ReliefF (RF)

ReliefF [26] is a supervised multivariate feature selection algorithm of the filter model which is the extension of Relief is a univariate model. Assuming that p instances are randomly sampled from data, the evaluation criterion for handling multiclass problems is of the form

$$\mathrm{SC}_R(\mathbf{f}_i) = \frac{1}{p} \cdot \sum_{t=1}^{p} \left\{ -\frac{1}{m_{\mathbf{x}_t}} \sum_{\mathbf{x}_j \in NH(\mathbf{x}_t)} d(f_{t,i} - f_{j,i}) + \sum_{y \neq y_{\mathbf{x}_t}} \frac{1}{m_{\mathbf{x}_t,y}} \frac{P(y)}{1 - P(y_{\mathbf{x}_t})} \sum_{\mathbf{x}_j \in NM(\mathbf{x}_t,y)} d(f_{t,i} - f_{j,i}) \right\}$$

where y_{xt} is the class label of the instance x_t and P(y) is the probability of an instance being from the class y. NH(x) or NM(x, y) denotes a set of nearest points to x with the same class of x, or a different class (the class y), respectively. m_{xt} and $m_{xt,y}$ are the sizes of the sets NH(x_t) and NM(x_t, y), respectively. Usually, the size of both NH(x) and NM(x, y); ¥ $y \neq y_{xt}$, is set to a pre-specified constant k.

B. Information Gain (IG)

Information Gain [6] is supervised univariate feature selection algorithm of the filter model which is a measure of dependence between the feature and the class label. It is one of the most powerful feature selection techniques and it is easy to compute and simple to interpret. Information Gain (IG) of a feature X and the class labels Y is calculated as

$$IG(X,Y) = H(X) - H(X|Y)$$

Entropy (H) is a measure of the uncertainty associated with a random variable. H(X) and H(X/Y) is the entropy of X and the entropy of X after observing Y, respectively.

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)).$$

The maximum value of information gain is 1. A feature with a high information gain is relevant. Information gain is evaluated independently for each feature and the features with the top-k values are selected as the relevant features. This feature selection algorithm does not eliminate redundant features.

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j))$$

C. Gain Ratio

The Gain Ratio is the non-symmetrical measure that is introduced to compensate for the bias of the IG [31]. GR is given by

$$GR = \frac{IG}{H(X)}$$

As the above equation presents, when the variable Y has to be predicted, the Information Gain has to normalized by dividing by the entropy of X, and vice versa. Due to this normalization, the Gain Ratio values always fall in the range [0, 1]. A value of Gain Ratio = 1 indicates that the knowledge of X completely predicts Y, and Gain Ratio = 0 means that there is no relation between Y and X. The Gain Ratio works well variables with fewer values where as the Information Gain works well variables with larger values.

D. Gini Index (GI)

Gini index [16] is supervised multivariate feature selection algorithm of the filter model to measure for quantifying a feature's ability to distinguish between classes. Given C classes, Gini Index of a feature *f* can be calculated as Gini Index can take the maximum value of 0.5 for a binary classification. The more relevant features have smaller Gini index values. Gini Index of each feature is calculated independently and the top k features with the smallest Gini index are selected. Like Information gain, it also not eliminates redundant features.

$$GiniIndex(f) = 1 - \sum_{i=1}^{C}[p(i|f)]^2$$

E. Random Forest (RF)

Random Forest developed by Leo Breiman [4] is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected using the induction process. Prediction is made by aggregating the predictions of the ensemble. Random Forest generally proves a significant performance improvement as compared to single tree classifier C4.5.

VI. EXPERIMENTAL RESULTS

The dimensionality of data used in research domain is rapidly increasing at many folds. The datasets may ranges from hundreds to more than thousands of features specifically in the field like genomic microarray analysis. Therefore, data reduction or dimensionality reduction is come to existence in order to improve the clustering or classification accuracy and throughput. This paper uses lung-cancer data set and this has given to all 5 feature selection algorithms. The dataset consists of 32 samples and each sample has 56 features. Not all 56 features are equally important and some may be less important which may decrease the performance and accuracy of the above said data mining tasks. This paper identified 10 features as most important features which are predominant in lung cancer dataset. This is not understood that other 46 features are not important, these are attribute may have comparatively less importance than the 10 predominant features. The predominant 10 features out of 56 features and results of various feature selection algorithms are shown in table 1.

Table 1 Top 10 Predominant Features with its Ranks

Attribute	RF	IG	GR	GI	SVM W	RF
a23	0.04	0.26	0.32	0.05	0.06	2.29
a56	0.16	0.24	0.28	0.04	0.06	1.18
a6	0.28	0.38	0.25	0.07	0.17	1.01
a20	0.36	0.47	0.35	0.08	0.07	0.96
a27	0.05	0.11	0.12	0.02	0.04	0.7
a19	0.26	0.38	0.4	0.08	0.02	0.59
a15	0.03	0.18	0.12	0.04	0.07	0.56
a21	0.06	0.04	0.05	0.01	0.01	0.42
a13	0.04	0.26	0.17	0.04	0.05	0.39
a53	0.06	0.21	0.24	0.04	0.04	0.33

This work also observes that RF outperforms than all other feature selection algorithms, IG and GR are the next better feature selection algorithms. The RF feature selection algorithm gave the next better result and the GI and SVMW gave equally poor performance in feature selection and ranking. The performance of all 5 filter feature selection algorithms is depicted in Figure 2.

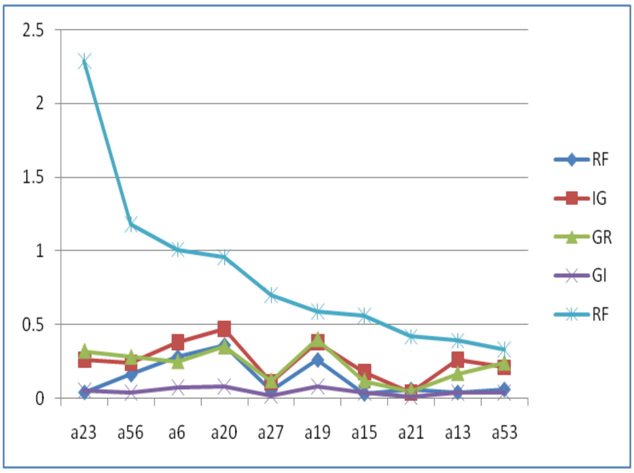

Fig 2 Performance of Feature selection Algorithms

VI. CONCLUSION

Feature selection has been a reported as ever green research topic with practical significance in many areas such as statistics, pattern recognition, machine learning, and data mining, web mining, text mining, image processing, and gene microarrays analysis. These feature selection algorithms are very well useful to build simpler and more comprehensible models, improving data mining tasks performance and accuracy, and helps to understand predominant data. This paper presents the analysis of all 5 feature selection algorithms and tested all algorithms by inputting lung cancer dataset. This work also presents that RF outperforms than all other feature selection algorithms, IG and GR are the next better feature selection algorithms. The RF feature selection algorithm gave the next better result and the GI gave equally poor performance in feature selection and ranking.

REFERENCES

[1] Han & Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2006.

[2] Huan Liu and Lei Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering.
IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, Pages: 491 - 502, 2005

[3] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis: An International Journal*, 1(3):131–156, 1997.

[4] P. Langley. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall Symposium on Relevance*, pages 140–144, 1994.

[5] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering – a filter solution. In *Proceedings of the Second International Conference on Data Mining*, pages 115–122, 2002.

[6] M.A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366, 2000.

[7] H. Liu and R. Setiono. A probabilistic approach to feature selection - a filter solution. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 319– 327, 1996.

[8] R. Caruana and D. Freitag. Greedy attribute selection. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 28–36, 1994.

[9] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 247–254, 2000.

[10] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 365–369, 2000.

[11] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 74–81, 2001.

[12] A. Y. Ng. On feature selection: learning with exponentially many irrelevant features as training examples. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 404–412, 1998.

[13] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 601–608, 2001.

[14] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[15] Inaki Inza, Pedro Larranaga, Rosa Blanco, and Antonio J. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. Arti_cial Intelligence in Medicine, 31:91-103, 2004.

[16] George Forman. An extensive empirical study of feature selection metrics for text classi_cation. Journal of Machine Learning Research, 3:1289-1305, 2003.

[17] R. Gonzalez and R. Woods. Digital Image Processing. Addison-Wesley, 2nd edition, 1993.

[18] W. Lee, S. J. Stolfo, and K. W. Mok. Adaptive intrusion detection: A data mining approach. AI Review, 14(6):533-567, 2000.

[19]A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets , Carmen Lai, Marcel JT Reinders, Laura J van't Veer and Lodewyk FA Wessels, BMC Bioinformatics 2006, 7:235 doi:10.1186/1471-2105-7-235

[20] Medical datamining with a new algorithm for Feature Selection and Naïve Bayesian classifier, Ranjit Abraham, Jay B.Simha, Iyengar S.S, 10th International Conference on Information Technology, 2007.

[21] Evaluating feature selection methods for learning in data mining applications, Selwyn Piramuthu, European Journal of Operational Research 156 (2004) 483–494.

[22] Rough set methods in feature selection and recognition, Roman W. Swiniarski a,, Andrzej Skowron b, Pattern Recognition Letters 24 (2003) 833–849, Elsevier Science.

[23] I. Kononenko. Estimating attributes : Analysis and extension of RELIEF. In F. Bergadano and L. De Raedt, editors, Proceedings of the European Conference on Machine Learning, April 6-8, pages 171-182, Catania, Italy, 1994. Berlin: Springer-Verlag.

[24] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, 1991.

[25] C. Gini. Variabilite e mutabilita. Memorie di metodologia statistica, 1912.

[26] Hall, M.A., and Smith, L.A., "Practical feature subset selection for machine learning", Proceedings of the 21st Australian Computer Science Conference, 1998, 181–191.

[27] Breiman, L., Random Forests, Machine Learning 45(1), 5-32, 2001.